Convención Internacional 2017 Universidad Central "Marta Abreu" de Las Villas

Conferencia Internacional de Procesamiento de la Información

Minería de datos educativos para la predicción personalizada del rendimiento académico
(Educational Data Mining for Personalized Prediction of Academic Performance)

J. del Campo-Ávila [a] G. Ramos-Jiménez [a] R. Morales-Bueno [a] M. Baena-García [b]

[a] *Universidad de Málaga, Andalucía Tech, Complejo Tecnológico, Campus de Teatinos, 29071 Málaga, España*

[b] *Vithas Salud Rincón, 29740 Torre del Mar (Málaga), España*

Resumen

La Minería de Datos Educativos (Educational Data Mining - EDM) está adquiriendo gran importancia como un nuevo campo de investigación interdisciplinario relacionado con algunas otras áreas. Está directamente relacionado con los Sistemas Educativos basados en la Web (Web-based Educational Systems - WBES) y la Minería de Datos (Data Mining - DM), siendo esta última una parte fundamental del Descubrimiento de Conocimiento en Bases de Datos (Knowledge Discovery in Databases - KDD).

Los WBES almacenan y administran grandes cantidades de datos. Estos datos están creciendo cada vez más y contienen conocimientos ocultos que podrían ser muy útiles para los usuarios (tanto profesores como estudiantes). Es conveniente identificar tales conocimientos en forma de modelos, patrones o cualquier otro esquema de representación que permita una mejor explotación del sistema. La minería de datos se revela como la herramienta para lograr tal descubrimiento, dando lugar a la EDM. En este contexto complejo se suelen utilizar distintas técnicas y algoritmos de aprendizaje para obtener los mejores resultados.

En este trabajo se estudia, para una asignatura de Informática Teórica, concretamente la asignatura "Teoría de Autómatas y Lenguajes Formales", cómo predecir el rendimiento académico alcanzado por los estudiantes, a partir de la realización de controles intermedios. Para ello se han aplicado y comparado distintos tipos de algoritmos de aprendizaje (vecinos más cercanos, árboles de decisión, multiclasificadores). Todo el proceso de control y evaluación de los estudiantes durante el curso se ha llevado a cabo a través de la herramienta web denominada SIETTE, desarrollada en nuestro departamento, y que además se utiliza en ámbitos fuera de nuestra propia universidad.

Palabras clave: Descubrimiento de Conocimiento, Minería de Datos, Sistemas Educativos Basados en la Web, Árboles de Decisión, Multiclasificadores

Email addresses: jcampo@uma.es (J. del Campo-Ávila), ramos@lcc.uma.es (G. Ramos-Jiménez),

.

1. Introducción

Desde que Internet abrió nuevas formas de comunicación, el sector educativo adoptó tal tecnología y desarrolló los Sistemas Educativos basados en la Web (WBES). Al principio eran sistemas estáticos, principalmente dedicados a la divulgación de contenidos. Pero progresivamente fueron ampliando sus características para hacerlos sistemas adaptativos e inteligentes [1]. En estos momentos existen muchos sistemas diferentes que combinan diversos elementos para alcanzar cierto nivel de inteligencia. Así, podemos encontrar WBES con técnicas adaptativas [2], otros WBES con mecanismos inteligentes [3] y sistemas más complejos que combinan ambas propiedades (una revisión detallada de tales sistemas fue presentada por Brusilovsky y Peylo [4]).

Lo que es evidente es el alto volumen de datos que estos sistemas almacenan y procesan continuamente: las relaciones entre los contenidos ofrecidos a los estudiantes, las interacciones con los estudiantes, el número de visitas, las puntuaciones obtenidas en las pruebas, el tiempo utilizado para responder a esas pruebas, etc.

El descubrimiento de conocimiento en bases de datos (KDD) continúa extendiéndose a casi todos los campos donde se almacenan y procesan grandes cantidades de datos (bases de datos, registros del sistema, registros de actividad, etc.), por lo que los WBES se convierten en otro entorno para aplicar procesos de KDD. Las técnicas de minería de datos son esenciales para uno de los puntos más importantes del KDD: se aplican en fase de análisis de datos y se utilizan algoritmos de aprendizaje automático para producir los modelos que resumen el conocimiento descubierto [5]. Por lo tanto, es fácil ver que las tareas educativas pueden beneficiarse del conocimiento extraído por la minería de datos.

Este campo de investigación se llama Minería de Datos Educativos (EDM) y su objetivo principal es analizar los datos almacenados en WBES para resolver problemas de investigación educativa [6]: validación del sistema educativo, predicción de logros de aprendizaje de los estudiantes, identificación de conceptos erróneos [7], evaluación y retroalimentación a los autores de los cursos [8], etc.

En este trabajo se estudia, para la asignatura de Informática Teórica "Teoría de Autómatas y Lenguajes Formales", cómo predecir los resultados de aprendizaje alcanzados por los estudiantes, a partir de la realización de controles intermedios. Mediante la recopilación de las respuestas dadas a las preguntas en dichos controles y su posterior análisis, se pretende conocer el desempeño que alcanzarán los diferentes estudiantes en la citada asignatura. Para ello se han aplicado y comparado distintos tipos de algoritmos de aprendizaje (vecinos más cercanos, árboles de decisión, multiclasificadores).

Esta comunicación está organizada de la siguiente manera. En primer lugar, en la Sección 2 se delimitan los objetivos que se pretenden conseguir. A continuación, en la Sección 3 se describen los materiales utilizados y la metodología llevada a cabo. Básicamente, los materiales utilizados han sido los datos recogidos por SIETTE (herramienta basada en la Web para pruebas adaptativas [9]), y los algoritmos de aprendizaje necesarios para hacer el análisis, disponibles en el entorno de trabajo para minería de datos llamado Weka [10]. Luego, en la Sección 4, presentamos la experimentación y los resultados obtenidos, comentando los patrones descubiertos por los algoritmos de aprendizaje automático. Finalmente, en la

rmorales@uma.es (R. Morales-Bueno), BaenaGM@vithas.es (M. Baena-García).

Sección 5, resumimos las conclusiones más relevantes y proponemos como línea de trabajo futura la realización de un estudio de campo, utilizando el modelo predictivo obtenido, para ver si la realimentación de las predicciones personalizadas a los estudiantes pueden mejorar sus resultados académicos.

2. Objetivos

En este trabajo se pretende estudiar la posibilidad de comprender y predecir el rendimiento que alcanzan los estudiantes en una asignatura concreta, partiendo de una serie de evidencias intermedias, recopiladas en forma de controles basados en preguntas de tipo test. La asignatura escogida para estudiar la viabilidad está en el ámbito de la Informática Teórica, concretamente la asignatura de "Teoría de Autómatas y Lenguajes Formales". Es una asignatura del segundo curso del Grado de Ingeniería Informática, que se imparte en la Escuela Técnica Superior de Ingeniería Informática de la Universidad de Málaga (España).

En esta asignatura se presentan diferentes tipos de conceptos, según se atienda a diversos criterios. Entre ellos se podría destacar la separación existente entre conceptos teóricos y conceptos procedimentales. Al mismo tiempo, también se pueden considerar diferencias atendiendo a la novedad del concepto, puesto que hay conceptos nuevos, que a su vez se basan en conceptos que deberían haber sido adquiridos previamente (en otras asignaturas o incluso en otros niveles educativos). Es conveniente comentar en este punto dichas diferenciaciones, porque estarán directamente relacionadas con las características de las propias preguntas que se incluirán en los controles de tipo test. Y, como se reflejará en los resultados, puede ser un criterio determinante para predecir el rendimiento que alcanzará un estudiante determinado.

La evaluación de esta asignatura se divide de la siguiente forma: 50 % de la nota para controles intermedios y 50 % de la nota para el control final (realizado al final del semestre). Se realizan 3 controles intermedios y cada uno de ellos tiene el mismo peso, por lo que la contribución a la nota de la asignatura de cada uno de ellos es de $1/6$. Como se ha comentado, se quiere estudiar el desempeño del alumnado considerando la información que ofrecen los controles intermedios, por lo que surgen dos objetivos que se pueden afrontar al mismo tiempo:

- determinar el resultado que se obtendrá en el control final (que supone el 50 % de la asignatura) y
- determinar el aprovechamiento global del estudiante (la asignatura se supera si se obtiene un cinco o más sobre diez al realizar la suma ponderada de las notas de todos los controles).

3. Metodología

Para el estudio hemos recopilado la información de diferentes grupos y cursos (entre 2014-2015 y 2015-2016), obteniendo un total de 124 estudiantes. El criterio para escoger dichos datos ha sido el tener asegurado que todos esos estudiantes han seguido el mismo proceso de evaluación, realizando los mismos controles bajo condiciones similares.

Cada uno de los controles intermedios se conforma a partir de una base de datos con preguntas agrupadas

según diferentes características relacionadas con la temática a evaluar. Tener dicha base de datos de preguntas permite que el repositorio de preguntas vaya aumentando y se vaya actualizando. Para cada uno de los controles intermedios realizados en el marco de este trabajo, el banco de preguntas de cada control consta de 30 preguntas, de las cuales se proponen a cada estudiante 20 preguntas elegidas de forma aleatoria. Por otro lado, el control final se confecciona con un total de 50 preguntas, seleccionadas manualmente del banco de preguntas.

Cada una de las preguntas presenta tres opciones, entre las cuales sólo hay una correcta. Como corresponde, desde el punto de vista de la teoría para test de respuestas cerradas (con tres opciones), cada pregunta erróneamente contestada resta la mitad de lo que suma cada una contestada correctamente, no sumando ni restando las dejadas en blanco.

3.1. Conjuntos de datos

Los conjuntos de datos que se han generado para realizar el proceso de minería de datos, han sido múltiples y variados. De esta forma se pueden analizar diferentes opciones y determinar si hay algún tipo de información que sea más relevante. En este subapartado se presentan los detalles más destacados para la creación de dichos conjuntos de datos.

Para cada estudiante se tiene registro sobre su respuesta a cada pregunta de cada uno de los controles intermedios, así como la nota que sacó en dichos controles. Además, dicha información podemos utilizarla agrupándola de diferentes formas. A continuación detallamos la información disponible y cómo se ha representado:

- respuesta dada a cada una de las preguntas (de cada uno de los controles intermedios). Si el estudiante no se presentó al control, en todas las preguntas de dicho control aparece el valor *"Ausente"*. Por otro lado, si una pregunta concreta no fue elegida en la composición del test de ese estudiante (recordar que se eligen 20 de 30 posibles), aparece el valor *"NoSale"*. En el caso de que la pregunta se le presentase y no se contestase, el valor en ese caso es *"Blanco"*. Por último, en el caso de que sí se contestase la pregunta existen dos opciones para ser estudiadas: a) respuesta concreta señalada (A, B o C) y b) acierto de la pregunta (sí o no).
  De esta forma, para cada una de las preguntas existen cinco o seis valores posibles (según se utilice la respuesta concreta o el acierto a la pregunta).
  La inclusión del valor *"NoSale"* permite seguir ampliando la base de datos con preguntas, porque se podrán contemplar muchas más combinaciones y se podrán fusionar los datos de diferentes estudiantes que no hayan respondido exactamente la misma configuración de preguntas. De no haber introducido dicho valor, se estaría obligando a configurar todos los controles intermedios de forma estática, dejando así de aprovechar la opción de configurar aleatoriamente las preguntas de cada estudiante.
- nota numérica obtenida (en cada uno de los controles intermedios). Utilizada como número real.

La información anterior se ve complementada con el resultado final alcanzado por el estudiante y que, como se explicó anteriormente, se convierte en nuestro objetivo. Así, los atributos de clase que estudiaremos son dos:

- nota obtenida en el control final (50 % de la calificación final).
- calificación final alcanzada en la asignatura (suma ponderada de los controles intermedios y final)

Las notas (y calificaciones) han sido categorizadas atendiendo a dos criterios alternativos no excluyentes:

- las calificaciones clásicas (*"suspenso"*, *"aprobado"*, *"notable"*, *"sobresaliente"*) más *"no presentado"*, lo que nos da cinco valores posibles.
- una clasificación más sencilla atendiendo a si *"supera"* o *"no supera"* la nota mínima de cinco para aprobar, y el *"no presentado"*, lo que nos da tres valores.

Para estudiar la posible influencia de la incorporación de nueva información, conforme los controles intermedios se iban realizando, se han creado diferentes conjuntos de datos. En los más simples únicamente se utilizaba la información del primer control intermedio, y progresivamente se ha añadido el resto de información (los dos primeros controles y los tres controles).

3.2. Algoritmos de aprendizaje

Existen múltiples métodos para realizar el aprendizaje automático en el ámbito de la clasificación. A continuación enumeramos y resumimos los métodos que vamos a utilizar:

- *vecinos más cercanos*: los algoritmos basados en instancias constituyen un enfoque bastante sencillo, pero que no tienen que corresponderse con un pobre resultado. Su estrategia es determinar qué instancias del conjunto de datos son más parecidas a la nueva observación. Para realizar la predicción se tienen en cuenta las instancias más parecidas y una función de distancia. Uno de los métodos más conocidos es el de los k vecinos más cercanos (donde k hace referencia al número de "vecinos" a considerar) [11]. La implementación que está disponible en Weka se llama IBk.
- *árboles de decisión*: los algoritmos que inducen árboles de decisión son ampliamente usados y existen múltiples variantes. Algunas de sus características más destacadas son la capacidad de inducir modelos comprensibles (incluso por personas no experimentadas), y la habilidad para particionar el problema recursivamente de forma que se vaya simplificando su resolución (siguiendo estrategias de "divide y vencerás"). Uno de los algoritmos más conocidos es C4.5 [12], que permite trabajar con atributos tanto nominales como numéricos, valores desconocidos, ruido, etc. La implementación de C4.5 que está disponible en Weka se llama J48.
- *sistemas multiclasificadores*: estos sistemas se benefician de la idea de combinar diferentes modelos (generados con algún algoritmo básico). Así, realizando algún tipo de variación que permita obtener diferentes modelos a partir del mismo conjunto de datos, se puede enfatizar el aprendizaje en características que un único modelo no sería capaz de analizar con tanto detalle. Por lo general, son métodos que alcanzan una mayor precisión, son robustos al ruido y no producen sobreajuste, pero, como contrapartida, dificultan la posibilidad de ser comprendidos (un árbol de decisión sería más fácil de interpretar que un conjunto de árboles que son combinados con alguna función de votación). Dos de los métodos más utilizados son Bagging y Random Forest [13]. Random Forest[14] induce un conjunto de árboles de decisión seleccionando de forma aleatoria el conjunto de atributos que puede usarse en la inducción de cada uno de esos árboles. También existe una implementación de RandomForest disponible en Weka (lo nombraremos como RF).

Además de estos enfoques se ha considerado otro más básico, que podría verse como un árbol de decisión en el que no se ha llegado a producir ninguna expansión. Concretamente se llama "sin reglas" (el nombre en Weka es ZeroR) y únicamente ofrece el valor más repetido (la moda) cuando la clase es nominal (si fuese numérica devolvería la media aritmética). El objetivo de incorporar este algoritmo tan básico es establecerlo como punto de referencia, con el que comparar el resto de algoritmos. Sería de esperar que cualquier algoritmo (que realice una búsqueda más "inteligente") fuese capaz de mejorarlo.

4. Resultados

Dada las distintas categorizaciones de los atributos, las dos posibles clases a estudiar, los distintos algoritmos utilizados, y otras variantes, el número de experimentos que se han realizado es elevado, por lo que sólo comentaremos en este trabajo los resultados más significativos (por ejemplo, no se muestran los resultados en los que se considera el acierto o fallo de preguntas concretas de los test).

Entre las variantes que se muestran se han incluido aquellas que han analizado la inclusión (o no) de las notas de los controles intermedios. El motivo es que se suponía (con razón como se verá después) que ninguna pregunta concreta de los controles iba a poder competir en capacidad predictiva con la nota de los controles intermedios, ya que dichas notas son un resumen de muchas preguntas y además influyen directamente en la nota final de la asignatura. De ahí el hacer el estudio tanto contando con las notas de los controles intermedios como sin utilizarlas. En este último caso lo interesante no sería mejorar la predicción, sino detectar preguntas clave dentro de la comprensión de la asignatura, objetivo también importante.

A todo esto hay que añadir que se han construido conjuntos de datos de forma acumulativa, incorporando inicialmente sólo el primer control (con todas las variaciones ya comentadas), posteriormente el primero y el segundo, y finalmente los tres controles. Como ya se ha adelantado, por razones obvias de espacio no vamos a exponer aquí todos los resultados de los experimentos realizados, sino sólo un grupo en el que se encuentra el resultado predictivo más significativo, que incluye además, como veremos, conocimiento relevante respecto a la pregunta más significativa. En el Cuadro 1 se recopilan las diferentes configuraciones estudiadas cuando la clase a predecir es la nota del control final, mientras que en el Cuadro 2 se considera como clase la calificación final obtenida en la asignatura.

Para obtener dichos resultados se han repetido 10 validaciones cruzadas con 10 particiones y se han realizado test estadísticos (t-test pareado) con una confianza del 95 % para buscar cuáles son significativamente mejores (indicado con una "$\oplus$") o significativamente peores (indicado con un "$\ominus$") que los valores del algoritmo que hemos escogido como referencia (ZeroR).

En general, y en contra de lo que se podría esperar, la media de los algoritmos IBk y J48 suelen ser ligeramente peor que la de ZeroR, a pesar de que podría pensarse que este último es un algoritmo más "ingenuo" que los otros dos. La media de Random Forest sí es mejor que la de ZeroR, lo cual en este caso no es ninguna sorpresa al ser un algoritmo de tipo Bagging. Los multiclasificadores suelen tener mejores resultados que los clasificadores simples. De hecho, si miramos los 24 casos, vemos que en 14 (2+1+5+6) de ellos obtiene resultados significativamente mejores, y en ninguno hay resultados significativamente

Clase	Calificación (sobr., not., aprob., susp.)				Supera / No supera			
Conjunto de datos	ZeroR	IBk	J48	RF	ZeroR	IBk	J48	RF
Respuestas 1	62,19	48,68 $\ominus$	54,39 $\ominus$	62,01	62,12	55,63	52,69 $\ominus$	65,3
Respuestas 1 + Notas	62,19	48,22 $\ominus$	53,68 $\ominus$	62,19	62,12	55,63	53,1 $\ominus$	64,85
Respuestas 1+2	62,19	46,81 $\ominus$	53,46 $\ominus$	66,2 $\oplus$	62,12	50,67 $\ominus$	59,27	66,19
Respuestas 1+2 + Notas	62,19	45,69 $\ominus$	55,42	66,2 $\oplus$	62,12	51,44 $\ominus$	60,26	68,03 $\oplus$
Respuestas 1+2+3	62,19	50,56 $\ominus$	51,51 $\ominus$	65,47	62,12	57,29	56,52	65,68
Respuestas 1+2+3 + Notas	62,19	48,18 $\ominus$	52,46 $\ominus$	65,47	62,12	55,88	59,53	66,65
Valor medio	62,19	48,02	53,49	64,59	62,12	54,42	56,9	66,12
Diferencias ($\oplus$/ /$\ominus$)		(0/0/6)	(0/1/5)	(2/4/0)		(0/4/2)	(0/4/2)	(1/5/0)

Cuadro 1
Tabla de resultados de experimentación para predecir la nota del control final.

Clase	Calificación (sobr., not., aprob., susp.)				Supera / No supera			
Conjunto de datos	ZeroR	IBk	J48	RF	ZeroR	IBk	J48	RF
Respuestas 1	42,74	30,99 $\ominus$	43,72	46,79	50,82	48,99	55,6	61,57 $\oplus$
Respuestas 1 + Notas	42,74	26,66 $\ominus$	44,14	50,46 $\oplus$	50,82	49,69	60,05 $\oplus$	64,37 $\oplus$
Respuestas 1+2	42,74	31,22 $\ominus$	40,04	56,04 $\oplus$	50,82	51,4	54,47	67,96 $\oplus$
Respuestas 1+2 + Notas	42,74	33,8	50,34	57,24 $\oplus$	50,82	54,29	66,62 $\oplus$	**71,57** $\oplus$
Respuestas 1+2+3	42,74	35,97	42,18	54,25 $\oplus$	50,82	54,63	55,06	65,76 $\oplus$
Respuestas 1+2+3 + Notas	42,74	32,36 $\ominus$	47,28	53,08 $\oplus$	50,82	54,12	66,1 $\oplus$	68,69 $\oplus$
Valor medio	42,74	31,83	44,62	52,98	50,82	52,19	58,26	65,67
Diferencias ($\oplus$/ /$\ominus$)		(0/2/4)	(0/6/0)	(5/1/0)		(0/6/0)	(3/3/0)	(6/0/0)

Cuadro 2
Tabla de resultados de experimentación para predecir la calificación final en la asignatura.

peores.

También es curioso que no se obtienen los mejores resultados cuando se incluyen los tres controles intermedios, sino cuando sólo se tienen dos. Podría pensarse que cuanta más información mejor, pero puede ocurrir que al incluir los tres controles se produzca una elección inicial de atributos algo peor. Concretamente en nuestra experimentación todos los grupos que incluyen los tres controles tienen en su árbol como raíz la nota del control tres, lo que al parecer finalmente, cuando se desarrolla todo el árbol, no es tan bueno para predecir como si tenemos la nota del control uno en la raíz (que es justamente la que tiene el árbol obtenido para el conjunto que incluye la opción con sólo los dos primeros controles intermedios).

Otro resultado, esta vez esperado, es que la predicción mejora, aunque sea levemente, cuanto incluimos las notas de los controles que estemos considerando como atributos adicionales para la predicción.

Respecto a los resultados del conjunto concreto que tiene la mejor predicción (en negrita en el Cuadro 2) vemos que asciende a un $71,57\,\%$, lo cual no es mal resultado cuando nos referimos a un entorno educativo como es el caso [15]. Esto significa que aproximadamente a mitad de la asignatura, tras hacer los dos primeros controles, y aunque aún sólo tengamos 1/3 de la nota ponderada de cada estudiante, ya podríamos hacer con este modelo generado una predicción personalizada para cada estudiante con un acierto por encima del $70\,\%$ sobre cada uno de dichos estudiantes superará o no la asignatura al final.

Hay otro aspecto interesante, y es el conocimiento sobre las preguntas que obtenemos. Si bien Random Forest es, como hemos señalado, un algoritmo multiclasificador, por lo que no es fácil consultar directamente el conocimiento generado dentro del modelo, sí podemos consultar fácilmente el árbol de decisión generado por el algoritmo J48 para este conjunto de datos (subrayado en el cuadro). Si bien su precisión es menor ($66,62\,\%$) sigue siendo estadísticamente mejor que la referencia, y tiene la ventaja de poder consultar, como hemos comentado, el conocimiento generado. Es de destacar también el reducido tamaño del árbol con tan sólo catorce reglas (o ramas). En la Figura 1 se presenta dicho árbol de decisión.

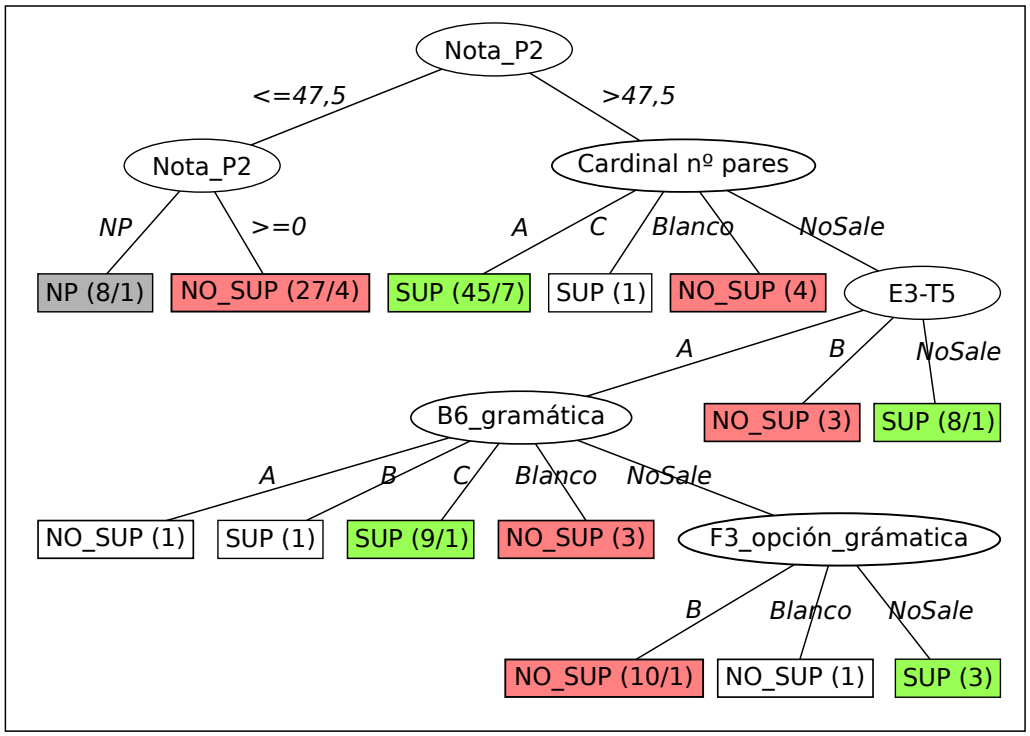

Figura 1. Árbol generado por el algoritmo J48 para cuando se quiere predecir la calificación del estudiante en la asignatura teniendo en cuenta la información de los dos primeros controles intermedios

Concretamente queremos destacar la primera pregunta de los controles que aparece, desde la raíz, al descender por el árbol: *"Cardinal n^o pares"*. Sin entrar en detalles de la asignatura, es una pregunta de fundamentos, de conceptos previos a la materia, ya que trata sobre los distintos cardinales infinitos. Esto muestra que la comprensión de un estudiante de los conceptos fundamentales de la asignatura son un buen indicador sobre el rendimiento final en la misma. También podría interpretarse como que es la parte de la asignatura que más difícil se le hace al estudiante entender, y que mejor discrimina por ello sobre su rendimiento final.

De hecho, esta hipótesis de que las preguntas de fundamentos son las más discriminantes a la hora de

predecir el rendimiento académico se confirma si miramos cuál es la siguiente pregunta en el árbol cuando la anterior no le ha aparecido al estudiante en el control (valor *"NoSale"*). La pregunta *"E3-T5"* es una pregunta sobre la denominada condición de bombeo regular, que también se puede considerar una pregunta de fundamentos, de un nivel de abstracción también alto.

Las otras dos preguntas que aparecen en el árbol, *"B6_gramática"* y *"F3_opción_gramática"*, son sobre las gramáticas. Concretamente la primera es sobre la relación entre gramáticas regulares izquierdas y derechas, y la segunda sobre las formas normales de las gramáticas regulares. Es decir, son preguntas del núcleo de la asignatura pero de corte más teórico que aplicado. Esto confirma la opinión general del profesorado de la asignatura de que a los estudiantes lo que más le cuesta de la misma son los aspectos teóricos, mientras que la aplicación de las gramáticas y sobre todo de los autómatas (destacar que no hay ninguna pregunta sobre ellos en el árbol) suele suponer menos esfuerzo para los estudiantes.

5. Conclusiones

Hay muchas más reflexiones que se podrían hacer sobre los datos obtenidos en esta experimentación, si bien con los resultados expuestos hemos cubierto el objetivo del trabajo que era mostrar la utilidad y potencial de la utilización de la minería de datos en entornos educativos, concretamente en la predicción del rendimiento académico dentro de una asignatura.

En el caso estudiado se ha visto que es posible predecir de manera personalizada, aproximadamente a mitad de la asignatura, y cuando aún sólo disponemos del 33 % de la calificación final del estudiante, si superará o no la misma con más del 70 % de acierto. El objetivo, entre otros, es que dicha información sirva para informar al estudiante y motivarlo para mejorar su rendimiento.

Además se ha detectado que la pregunta más predictiva es una de fundamentos matemáticos anterior al núcleo de la asignatura propiamente dicha. También la segunda pregunta más predictiva es de tipo fundamental. Este clase de descubrimientos puede ayudar a diseñar una mejor metodología docente, haciendo hincapié en las preguntas clave, de manera que el estudiante vea de forma más homogénea la dificultad de la asignatura.

Como trabajo futuro planteamos dos líneas. Por una parte, a nivel de algoritmos de aprendizaje, queremos no sólo probar con más algoritmos, sino también binarizar los datos para ver si, al igual que en otros contextos, dicha binarización mejora los índices de predicción. La segunda línea de trabajo futuro tiene que ver con el aspecto mencionado de dar información sobre las predicciones personalizadas a los estudiantes: informar a un grupo de estudiantes, al terminar el segundo test, de la predicción personalizada sobre si superará o no la asignatura si sigue en con mismo nivel de estudios; mientras que a otro grupo (que sería el grupo de control), sólo se le animaría a mejorar su nivel de estudios, pero sin la predicción personalizada. El objetivo sería ver si suministrar la predicción personalizada aumenta significativamente el porcentaje de estudiantes aprobados.

Por todo lo expuesto creemos que la minería de datos educativos para la predicción personalizada del rendimiento académico es un campo de investigación interesante, tanto a nivel teórico de aprendizaje

automático, como a nivel aplicado dentro del ámbito docente.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el I Plan Propio de Investigación y Transferencia de la Universidad de Málaga.

Referencias

[1] P. Brusilovsky, E. Schwarz, and G. Weber. Elm-art: An intelligent tutoring system on world wide web. *Lecture Notes in Computer Science*, 1086:261–269, 1996.

[2] C. Romero, S. Ventura, A. Zafra, and P. de Bra. Applying web usage mining for personalizing hyperlinks in web-based adaptive educational systems. *Computers & Education*, 53:828–840, 2009.

[3] E. Melis, E. Andrès, J. Büdenbender, A. Frishauf, G. Goguadse, P. Libbrecht, M. Pollet, and C. Ullrich. Activemath: A generic and adaptive web-based learning environment. *International Journal of Artificial Intelligence in Education*, 12:385–407, 2001.

[4] P. Brusilovsky and C. Peylo. Adaptive and intelligent web-based educational systems. *International Journal of Artificial Intelligence in Education*, 13:156–169, 2003.

[5] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17:37–54, 1996.

[6] C. Romero and S. Ventura. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics-Part C*, 40:601–618, 2010.

[7] E. Guzmán, R. Conejo, and J. Gálvez. A data-driven technique for misconception elicitation. *Lecture Notes in Computer Science*, 6075:243–254, 2010.

[8] C. Romero, S. Ventura, and P. De Bra. Knowledge discovery with genetic programming for providing feedback to courseware author. *User Model. User-Adapted Interaction*, 14:425–464, 2004.

[9] R. Conejo, E. Guzmán, E. Millán, M. Trella, J. L. Pérez-De-La-Cruz, and A. Ríos. Siette: A web-based tool for adaptive testing. *International Journal of Artificial Intelligence in Education*, 14: 29–61, 2004.

[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *SIGKDD Explorations Newsletter*, 11(1):10–18, 2009. ISSN 1931-0145. doi: 10.1145/1656274.1656278. URL `http://doi.acm.org/10.1145/1656274.1656278`.

[11] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21– 27, 1967.

[12] J. R. Quinlan. *C4.5: Programs for machine learning*. Morgan Kaufmann, 1993.

[13] A. J. Wyner, M. Olson, J. Bleich, and D. Mease. Explaining the success of adaboost and random forests as interpolating classifiers. *Journal of Machine Learning Research*, 18(48):1–33, 2017.

[14] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[15] S. T. Jishan, R. I. Rashu, N. Haque, and R. M. Rahman. Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. *Decision Analytics*, 2(1), 2015.